Proof-theoretic Analysis of Rationality for Strategic Games with Arbitrary Strategy Sets

Jonathan A. Zvesper[1] and Krzysztof R. Apt[2,3]

[1] Oxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, UK
[2] Centre for Mathematics and Computer Science (CWI), Science Park 123, 1098 XG Amsterdam, the Netherlands
[3] University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands

Abstract. In the context of strategic games, we provide an axiomatic proof of the statement

(Imp) Common knowledge of rationality implies that the players will choose only strategies that survive the iterated elimination of strictly dominated strategies.

Rationality here means playing only strategies one believes to be best responses. This involves looking at two formal languages. One, $\mathcal{L}_O$, is first-order, and is used to formalise optimality conditions, like avoiding strictly dominated strategies, or playing a best response. The other, $\mathcal{L}_\nu$, is a modal fixpoint language with expressions for optimality, rationality and belief. Fixpoints are used to form expressions for common belief and for iterated elimination of non-optimal strategies.

1 Introduction

There are two main sorts of solution concepts for strategic games: *equilibrium* concepts and what might be called "*effective*" concepts. One interpretation of the equilibrium concepts, for example Nash equilibrium, tacitly presupposes that a game is played repeatedly (see, e.g. [13, page 14]). Thus the standard condition for Nash equilibrium in terms of the knowledge or beliefs of the players [3] – the so-called "epistemic analysis" of Nash equilibrium – includes a requirement that players know the other players' strategy choices.

	L	R
L	$1,1$	$0,0$
R	$0,0$	$1,1$

	L	R
U	$1,1$	$1,0$
D	$0,0$	$0,1$

Fig. 1. Two strategic games

Consider the left-hand game in Figure 1, in which each player has two choices L and R and both players get payoff of 1 if they coordinate, and 0 otherwise. Then there are two Nash equilibria[1]: both play L or both play R. But this does not translate by itself into an effective strategy for either player reasoning in isolation, without some exogenous information.

In contrast, *effective* solution concepts, for example the iterated elimination of strictly dominated strategies, are compatible with such a "one-shot" interpretation of the game. Thus the epistemic analysis of the iterated elimination of strictly dominated strategies does not require that the players know each other's strategy choice.

A strategy s_i is strictly dominated if there is an alternative strategy t_i such that no matter what the opponent does, t_i is (strictly) better for i than s_i. Say that a player is *sd*-rational if he never plays a strategy that he believes to be strictly dominated. What the iterated elimination of strictly dominated strategies does in general require, see [4], is then that players have *common true belief* that each other is rational, that is: they are rational, believe that all are rational, believe that all believe that all are rational, etc.

In the right-hand game in Figure 1, the column player, on first looking at her choices L or R is, superficially, in the same situation as before: choose L and risk the opponent playing D or choice R and risk the opponent playing U. However, this time the row player can immediately dismiss playing D on the grounds that U will *always* be better, no matter what the column player does. So if the column player knows (or believes) this, then he cannot *rationally* play R, and so must play L.

In this paper we study the logical form of epistemic characterisation results of this second kind, so we give formal proof-theoretic principles to justify some given effective or algorithmic process in terms of common belief of some form of rationality. We will introduce two formal languages. One, $\mathcal{L}_O$, is a first-order language, that can be used to define 'optimality conditions'. Avoiding playing a strictly dominated strategy is an example of an 'optimality condition'. Another one is choosing a best response.

However, as observed in [2] for all such notions there are two versions: 'local' and 'global'. Notice that in our informal description of when s_i is strictly dominated by t_i we did not specify *where* i is allowed to choose alternative strategies from. In particular, since we are thinking of an iterated procedure, if t_i has been eliminated already then it would seem unreasonable to say that i should consider it. That intuition yields the

[1] A Nash equilibrium in a two-player game is a pair (s_1, s_2) of strategies, one for each player such that s_1 is a best response to s_2 and vice-versa.

local definition; the *global* definition states the opposite: that player i should always consider his original strategy set from the full game when looking to see if a strategy is dominated.

A motivation for looking at global versions of optimality notions is that they are often mathematically better behaved. On finite games the iterations for various local and global versions coincide [1], but on infinite games they can differ. In a nutshell: an optimality condition ϕ_i for player i is *global* if i does not 'forget', during the iterated elimination process, what strategies he has available in the whole game. The distinction is clarified in the respective definitions in $\mathcal{L}_O$.

An optimality condition ϕ induces an optimality operator O_ϕ on the complete lattice of restrictions (roughly: the subgames) of a given game. Eliminating non-ϕ-optimal strategies can be seen as the calculation of a fixpoint of the corresponding operator O_ϕ. Furthermore, common belief is characterised as a fixpoint (cf. Note 3 below). Viewed from the appropriate level of abstraction, in terms of fixpoints of operators, this connection between common belief of rationality and the iterated elimination of non-optimal strategies becomes clear.

We define a language $\mathcal{L}_\nu$ that describes things from this higher level of abstraction. Each optimality condition defines a corresponding notion of *rationality*, which means playing a strategy that one *believes* to be ϕ-optimal. $\mathcal{L}_\nu$ is a modal fixpoint language with modalities for *belief* and *optimality*, and so can express connections between optimality, rationality and (common) belief.

We say that an operator O on an arbitrary lattice $(D, \subseteq)$ is ***monotonic*** when for all $A, B \in D$, if $A \subseteq B$ then $O(A) \subseteq O(B)$. The global versions of relevant optimality operators, in particular of the operators corresponding to the best response and strict dominance, are monotonic. This is immediately verifiable in $\mathcal{L}_O$ by observing that the relevant definition is *positive*.

Our first result is a syntactic proof of the following result, where ϕ is a *monotonic* optimality condition:[2]

Theorem 1 *Common true belief of ϕ-rationality entails all played strategies survive the iterated elimination of non-ϕ-optimal strategies.*

Although this theorem relies on a rule for fixpoint calculi that is only sound for monotonic operators, the semantics of the language $\mathcal{L}_\nu$ allows also for arbitrary ***contracting*** operators, i.e. such that for all A, $O(A) \subseteq$

[2] By "common true belief" we mean a common belief that is correct. In particular, common knowledge entails common true belief.

A. We are therefore able to look at what more is needed in order to justify the following statement (cf. [4, Proposition 3.10]), where *gbr*-rationality means avoiding avoiding strategies one believes to be never best responses in the global sense:

Theorem 2 (**Imp**) *Common true belief of gbr-rationality implies that the players will choose only strategies that survive the iterated elimination of strictly dominated strategies.*

This theorem connects a global notion of *gbr*-rationality with a local one, referred to in the iterated elimination operator. Our language allows for arbitrary contracting operators, and their fixpoints to be formed, and we exhibit one sound rule connecting the resulting fixpoints with monotonic fixpoints.

Our theorems hold for arbitrary games, and the resulting potentially transfinite iterations of the elimination process. The syntactic approach clarifies the logical underpinnings of the epistemic analysis. It shows that the use of transfinite iterations can be naturally captured in $\mathcal{L}_\nu$, at least when the relevant operators are monotonic, by a single inference rule that involves greatest fixpoints.

The relevance of monotonicity in the context of epistemic analysis of finite strategic games has already been pointed out in [5], where the connection is also noted between the iterated elimination of non-optimal strategies and the calculation of the fixpoint of the corresponding operator.

To our knowledge, although several languages have been suggested for reasoning about strategic games (e.g. [7]), none use explicit fixpoints (except, as we mentioned, for some suggestions in [5]) and none use arbitrary optimality operators.

Therefore they are not appropriate for reasoning at the level of abstraction that we suggest when studying the epistemic foundations of these "effective" solution concepts. For example while [7, Section 13] does provide some analysis of the logical form of the argument that common knowledge of one kind of rationality implies not playing strategies that are strictly dominated, the fixpoint reasoning is done at the meta-level. What [7] provides is a proof schema, that shows how, for any finite game, and any natural number n, to give a proof that common knowledge of rationality entails not playing strategies that are eliminated in n rounds of elimination of non-optimal strategies.

The more general and elegant reasoning principle is captured by using fixpoint operators and optimality operators. Another important advan-

tage to our approach is that we are not restricted in our analysis to finite games. This means in particular that our logical analysis covers the mixed extension of any finite game.

Our use of transfinite iterations is motivated by the original finding of [12], where a two-player game is constructed for which the ω_0 (the first infinite ordinal) and $\omega_0 + 1$ iterations of the rationalizability operator of [6] differ.

2 Games and the language $\mathcal{L}_O$

A ***strategic game*** is a tuple $(T_1, \ldots, T_n, <_1, \ldots, <_n)$, where $\{1, \ldots, n\}$ are the players and each T_i is player i's set of strategies, and $<_i$ is player i's preference relation, which is a total linear order over the set of ***strategy profiles*** $T = \prod_{i=1}^{n} T_i$. Note that we assume arbitrary games, rather than restricting to games in which T is finite. To depict games it is sometimes easier, as we did in Figure 1, to write down a number for the players' "payoffs", rather than just a preference ordering. We use some standard notation from game theory, writing s_{-i} for $(s_1, \ldots s_{i-1}, s_{i+1}, \ldots s_n)$ and (s_i, t_{-i}) for the strategy profile $(t_1, \ldots t_{i-1}, s_i, t_{i+1}, \ldots s_n)$, as well as S_{-i} for $\prod_{j \neq i} S_j$. A ***restriction*** of the game $(T_1, \ldots, T_n, <_1, \ldots, <_n)$ is a sequence $S = (S_1, \ldots, S_n)$ with $S_i \subseteq T_i$ for all players i, i.e. a (possibly empty) subgame in which the payoff information is left out.

The language we use for specifying optimality conditions is a first-order language, with variables $V = \{x, y, z, \ldots\}$, a monadic predicate C, a constant o and a family of n ternary relation symbols $\cdot \geq^i \cdot$, where $i \in [1..n]$. So $\mathcal{L}_O$ is given by the following inductive definition:

$$\phi ::= C(a) \mid a \geq_c^i b \mid \neg\phi \mid \phi \wedge \phi \mid \exists x \phi,$$

where $i \in [1..n]$ and $\{a, b, c\} \subseteq V \cup \{o\}$.

We use the standard abbreviations $\rightarrow$ and $\vee$, further abbreviate $\neg a \geq_c^i b$ to $b >_c^i a$, $\forall x\phi$ to $\neg\exists x\neg\phi$, $\exists x(C(x) \wedge \phi)$ to $\exists x \in C\, \phi$, and $\forall x(C(x) \rightarrow \phi)$ to $\forall x \in C\, \phi$.

An *optimality model* (G, G', s) is a triple consisting of a strategic game $G = (T_1, \ldots, T_n, <_1, \ldots, <_n)$, a restriction G' of G, and a strategy profile $s \in T$. G will be used to interpret the predicate C, and s will be the interpretation of o. An *assignment* for (G, G', s) is a function α assigning a strategy profile in T to each variable, and s to o. The ternary satisfaction relation $\models$ between optimality models, assignments and formulas of $\mathcal{L}_O$ is defined inductively as follows, where α is an assignment for (G, G', s),

and $\not\models$ the complement of $\models$:

$$\begin{array}{ll}
(G, G', s) \models_\alpha C(x) & \Leftrightarrow \forall i \in \{1, \ldots, n\},\ (\alpha(x))_i \in G'_i \\
(G, G', s) \models_\alpha x \geq^i_z y & \Leftrightarrow (\alpha(x)_i, \alpha(z)_{-i}) \geq_i (\alpha(y)_i, \alpha(z)_{-i}) \\
(G, G', s) \models_\alpha \neg\phi & \Leftrightarrow (G, G', s) \not\models_\alpha \phi \\
(G, G', s) \models_\alpha \phi_1 \wedge \phi_2 & \Leftrightarrow (G, G', s) \models_\alpha \phi_1 \text{ and } (G, G', s) \models_\alpha \phi_2 \\
(G, G', s) \models_\alpha \exists x\phi & \Leftrightarrow \text{there is } \alpha' : (G, G', s) \models_{\alpha'} \phi \text{ and} \\
 & \qquad \forall y \in V \text{ with } x \neq y,\ \alpha(y) = \alpha'(y)
\end{array}$$

If for any assignment α for G we have $(G, G', s) \models_\alpha \phi$ then we write $(G, G', s) \models \phi$. A variable x occurs ***free*** in ϕ if it is not under the scope of a quantifier $\exists x$; a formula is ***closed*** if it has no free variables.

An ***optimality condition*** for player i is a closed $\mathcal{L}_O$-formula in which all the occurrences of the atomic formulas $a \geq^j_c b$ are with j equal to i. Intuitively, an optimality condition ϕ_i for player i is a way of specifying what it means for i's strategy in o to be an 'OK' choice for i given that i's opponents will play according to C_{-i} and that i's alternatives are C_i.

In particular, we are interested in the following optimality conditions:

- $lsd_i := \forall y \in C\, \exists z \in C\, o \geq^i_z y$,
- $gsd_i := \forall y\, \exists z \in C\, o \geq^i_z y$,
- $gbr_i := \exists z \in C\, \forall y\, o \geq^i_z y$.

The optimality conditions listed define some fundamental notions from game theory: lsd_i says that o_i is not *locally* strictly dominated in the context of C; gsd_i says that o_i is not *globally* strictly dominated in the context of C; and gbr_i says that o_i is globally a best response in the context of C.

The distinction between local and global properties, studied further in [2], is clarified below. It important for us here because the global versions, in contrast to the local ones, satisfy a syntactic property to be defined shortly.

First, as an illustration of the difference between gbr_i and gsd_i, consider the game in Figure 2. Call that game H, with the row player 1 and the column player 2. Then we have

$$(H, (T_1, T_2), (D, R)) \models gsd_1,$$

but

$$(H, (T_1, T_2), (D, R)) \models \neg gbr_1.$$

	L	R
U	$2,1$	$0,0$
M	$0,1$	$2,0$
D	$1,0$	$1,2$

Fig. 2. An illustration of the difference between strict dominance and best response

The local notions are such that when the 'context' restriction C consists of a singleton strategy for a player i, then that strategy is locally optimal. So for example

$$(H, (\{U, M\}, \{R\}), (U, R)) \models lsd_2,$$

whereas

$$(H, (\{U, M\}, \{R\}), (U, R)) \models \neg gsd_2.$$

We say that an optimality condition ϕ_i is ***positive*** when any subformula of the form $C(z)$, with z any variable, occurs under the scope of an even number of negation signs ($\neg$). Note that both gbr_i and gsd_i are positive, while lsd_i is not. As we will see in a moment, positive optimality conditions induce monotonic optimality operators, and monotonicity will be the condition required of optimality operators in Theorem 1 relating common knowledge of ϕ-rationality with the iterated elimination of non-ϕ strategies.

3 Optimality operators

Henceforth let $G = (T_1, \ldots, T_n, <_1, \ldots, <_n)$ be a fixed strategic game. Recall that a *restriction* of the game G is a sequence $S = (S_1, \ldots, S_n)$ with $S_i \subseteq T_i$ for all players i. We will interpret optimality conditions as *operators* on the lattice of the restrictions of a game ordered by component-wise set inclusion:

$$(S_1, \ldots, S_n) \subseteq (S'_1, \ldots, S'_n) \text{ iff } S_i \subseteq S'_i \text{ for all } i \in [1..n].$$

Given a sequence ϕ giving an optimality condition ϕ_i for each player i, we introduce an ***optimality operator*** O_ϕ defined by

$$O_\phi(S) = \prod_{i=1}^{n} \{s_i \in S_i \mid \phi_i(s_i, S).\}$$

Consider now an operator O on an arbitrary complete lattice $(D, \subseteq)$ with largest element $\top$. We say that an element $S \in D$ is a ***fixpoint*** of O if $S = O(S)$ and a ***post-fixpoint*** of O if $S \subseteq O(S)$.

We define by transfinite induction a sequence of elements O^α of D, for all ordinals α:

- $O^0 := \top$,
- $O^{\alpha+1} := O(O^\alpha)$,
- for limit ordinals β, $O^\beta := \bigcap_{\alpha<\beta} O^\alpha$.

We call the least α such that $O^{\alpha+1} = O^\alpha$ the ***closure ordinal*** of O and denote it by α_O. We call then O^{α_O} the ***outcome of*** (iterating) O and write it alternatively as O^∞.

Not all operators have fixpoints, but the monotonic and contracting ones (already defined in the introduction) do:

Note 1. Consider an operator O on $(D, \subseteq)$.

(i) If O is contracting or monotonic, then it has an outcome, i.e., O^∞ is well-defined.
(ii) The operator $\overline{O}$ defined by $\overline{O}(X) := O(X) \cap X$ is contracting.
(iii) If O is monotonic, then the outcomes of O and $\overline{O}$ coincide.

Proof. For (i), it is enough to know that for every set D there is an ordinal α such that there is no injective function from α to D.

Note that the operators O_ϕ are by definition contracting, and hence all have outcomes. Furthermore, it is straightforward to verify that if ϕ_i is positive for all players i, then O_ϕ is monotonic.

The following classic result due to [14] also forms the basis of the soundness of some part of the proof systems we consider.[3]

Tarski's Fixpoint Theorem For every monotonic operator O on $(D, \subseteq)$

$$O^\infty = \nu O = \cup\{S \in D \mid S \subseteq O(S)\},$$

where νO is the largest fixpoint of O.

We shall need the following lemma, which is crucial in connecting iterations of arbitrary contracting operators with those of monotonic operators. It also ensures the soundness of one of the proof rules we will introduce.

[3] We use here its 'dual' version in which the iterations start at the largest and not at the least element of a complete lattice.

Lemma 1. *Consider two operators O_1 and O_2 on $(D, \subseteq)$ such that*

- *for all $S \in D$, $O_1(S) \subseteq O_2(S)$,*
- *O_1 is monotonic.*

Then $O_1^\infty \subseteq \overline{O_2}^\infty$.

Proof. By Note 1(i) the outcomes of O_1 and $\overline{O_2}$ exist.

We prove now by transfinite induction that for all α

$$\overline{O_1}^\alpha \subseteq \overline{O_2}^\alpha$$

from which the claim follows, since by Note 1(iii) we have $O_1^\infty = \overline{O_2}^\infty$.

By the definition of the iterations we only need to consider the induction step for a successor ordinal. So suppose the claim holds for some α.

The second assumption implies that $\overline{O_1}$ is monotonic. We have the following string of inclusions and equalities, where the first inclusion holds by the induction hypothesis and monotonicity of $\overline{O_1}$ and the second one by the first assumption

$$\overline{O_1}^{\alpha+1} = \overline{O_1}(\overline{O_1}^\alpha) \subseteq \overline{O_1}(\overline{O_2}^\alpha) = O_1(\overline{O_2}^\alpha) \cap \overline{O_2}^\alpha \subseteq O_2(\overline{O_2}^\alpha) \cap \overline{O_2}^\alpha = \overline{O_2}^{\alpha+1}.$$

4 Beliefs and the modal fixpoint language $\mathcal{L}_\nu$

Recall that G is a game $(T_1, \ldots, T_n, P_1, \ldots, P_n)$. A ***belief model*** for G is a tuple $(\Omega, \overline{s}_1, \ldots, \overline{s}_n, P_1, \ldots, P_n)$, with Ω a non-empty set of 'states', and for each player i, $\overline{s}_i : \Omega \to T_i$ and $P_i : \Omega \to 2^\Omega$. The P_i's are *possibility correspondences* cf. [4]. The idea of a possibility correspondence P_i is that if the actual state is ω then $P_i(\omega)$ is the set of states that i considers possible: those that i considers might be the actual state.

Subsets of Ω are called ***events***. A player i *believes* an event E if that event holds in every state that i considers possible. Thus at the state ω, player i believes E iff $P_i(\omega) \subseteq E$.

Given some event E we write G_E to denote the restriction of G determined by E:

$$(G_E)_i = \{s_i \in T_i \mid \exists u \in E : \overline{s}_i(u) = s_i\}.$$

In the rest of this section we present a formal language $\mathcal{L}_\nu$ that will be interpreted over belief models. To begin, we consider the simpler language $\mathcal{L}$, the formulas of which are defined inductively as follows, where $i \in [1..n]$:

$$\psi ::= rat_{\phi_i} \mid \psi \wedge \psi \mid \neg\psi \mid \Box_i\psi \mid O_{\phi_i}\psi,$$

with ϕ_i an optimality condition for player i. We abbreviate the formula $\bigwedge_{i\in[1..n]} rat_{\phi_i}$ to rat_ϕ, $\bigwedge_{i\in[1..n]} \Box_i\psi$ to $\Box\psi$ and $\bigwedge_{i\in[1..n]} O_{\phi_i}\psi$ to $O_\phi\psi$.

Formulas of $\mathcal{L}$ are interpreted as events in (i.e. as subsets of the domain of) belief models. Given a belief model $(\Omega, \overline{s}_1, \ldots, \overline{s}_n, P_1, \ldots, P_n)$ for G, we define the ***interpretation function*** $[\![\cdot]\!] : \mathcal{L} \to \mathcal{P}(\Omega)$ as follows:

- $[\![rat_{\phi_i}]\!] = \{\omega \in \Omega \mid \phi_i(\overline{s}_i(\omega), G_{P_i(\omega)})\}$,
- $[\![\phi \wedge \psi]\!] = [\![\phi]\!] \cap [\![\psi]\!]$,
- $[\![\neg\psi]\!] = \Omega - [\![\psi]\!]$,
- $[\![\Box_i\psi]\!] = \{\omega \in \Omega \mid P_i(\omega) \subseteq [\![\psi]\!]\}$,
- $[\![O_{\phi_i}\psi]\!] = \{\omega \in \Omega \mid (G, G_{[\![\psi]\!]}, \overline{s}_i(\omega)) \models \phi_i\}$.

$P_i(\omega)$ gives the set of states that i considers possible at ω, so $[\![rat_{\phi_i}]\!]$ is the event that player i is ϕ_i-rational, since it means that i's strategy is optimal according to ϕ_i in the context that the player considers it possible that he is in. The semantic clause for $\Box_i$ was mentioned at the begin of this section and is familiar from epistemic logic: $[\![\Box_i\psi]\!]$ is the event that player i believes the event $[\![\psi]\!]$. $[\![O_{\phi_i}\psi]\!]$ is the event that player i's strategy is optimal according to the optimality condition ϕ_i, in the context of the restriction $G_{[\![\psi]\!]}$.

Then clearly $[\![rat_\phi]\!]$ is the event that every player i is ϕ_i-rational; $[\![O_\phi\psi]\!]$ is the event that every player's strategy is ϕ_i-optimal in the context of the restriction $G_{[\![\psi]\!]}$; and $[\![\Box\psi]\!]$ is the event that every player believes the event $[\![\psi]\!]$ to hold.

Although $\mathcal{L}$ can express some connections between our formal definitions of optimality rationality and beliefs, it could be made more expressive. The language could be extended with, for example, atoms s_i expressing the event that the strategy s_i is chosen. This choice is made for example in [7], where modal languages for reasoning about games are defined. The language we introduce is not parametrised by the game, and consequently can unproblematically be used to reason about games with arbitrary strategy sets.

We will use our language to talk about fixpoint notions: common belief and iterated elimination of non-optimal strategies. Let us therefore explain what is meant by ***common belief***. Common belief of an event E is the event that all players believe E, all players believe that they believe E, all players believe that they believe that…, and so on. Formally, we

define $\mathcal{CB}(E)$, the event that E is commonly believed, inductively:

$$\begin{aligned}\mathcal{B}_1(E) &= \{\omega \in \Omega \mid \forall i \in [1..n],\ P_i(\omega) \subseteq E\}\\ \mathcal{B}_{m+1}(E) &= \mathcal{B}_1(\mathcal{B}_m(E))\\ \mathcal{CB}(E) &= \bigcap_{m>0} \mathcal{B}_m(E)\end{aligned}$$

Notice that $\mathcal{B}_1(E)$ is the event that everybody believes that E (indeed, we have $\mathcal{B}_1[\![\psi]\!] = [\![\Box\psi]\!]$), $\mathcal{B}_2(E)$ is the event that everybody believes that everybody believes that E, etc.

'Common belief' is called 'common knowledge' when for all players i and all states $\omega \in \Omega$, we have $\omega \in P_i(\omega)$. In such a case the players have never ruled out the current state, and so it is legitimate to interpret $\Box_i\psi$ as 'i knows that ψ'.

Both common knowledge and common belief are known to have equivalent characterisations as fixpoints, and we will exploit this below in defining them in the modal fixpoint language which we now specify.

We extend the vocabulary of $\mathcal{L}$ with a single set variable denoted by X and the contracting fixpoint operator νX. (The corresponding extension of first-order logic by the dual, inflationary fixpoint operator μX was first studied in [8].) Modulo one caveat the resulting language $\mathcal{L}_\nu$ is defined as follows:

$$\psi ::= rat_{\phi_i} \mid (\psi \wedge \psi) \mid \neg\psi \mid \Box_i\psi \mid O_{\phi_i}\psi \mid \nu X.\psi$$

The caveat is the following:

- ϕ must be ***ν-free***, which means that it does not contain any occurrences of the νX operator.

This restriction is not necessary but simplifies matters and is sufficient for our considerations.

To extend the interpretation function $[\![\cdot]\!]$ to $\mathcal{L}_\nu$, we must keep track of the variable X. Therefore we first extend the function $[\![\cdot]\!] : \mathcal{L} \to \mathcal{P}(\Omega)$ to a function $[\![\cdot \mid \cdot]\!] : \mathcal{L}_\nu \times \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ by padding it with a dummy argument. We give one clause as an example:

- $[\![\Box_i\psi \mid E]\!] = \{\omega \in \Omega \mid P_i(\omega) \subseteq [\![\psi \mid E]\!]\}$.

We use this extra argument in the semantic clause for the variable X:

- $[\![X \mid E]\!] = E$.

Those formulas whose semantics we have so far given define operators. More specifically, for each of them $[\![\psi \mid \cdot]\!]$ is an operator on the powerset $\mathcal{P}(\Omega)$ of Ω. We use this to define the clause for νX:

– $[\![\nu X.\psi \mid E]\!] = ([\![\psi \wedge X \mid \cdot]\!])^{\infty}$.

When X does not occur free in ψ, we have $[\![\psi \mid E]\!] = [\![\psi \mid F]\!]$ for any events E and F, so in these cases we can write simply $[\![\psi]\!]$. Note that $[\![\nu X.\psi]\!]$ is well-defined since for all E we have $[\![\psi \wedge X \mid E]\!] = [\![\psi \mid E]\!] \cap [\![X \mid E]\!] \subseteq E$, so the operator $[\![\psi \wedge X \mid \cdot]\!]$ is contracting.

We say that a formula ψ of $\mathcal{L}_\nu$ is ***positive in*** X when each occurrence of X in ψ is under the scope of an even number of negation signs ($\neg$), and under the scope of an optimality operator O_{ϕ_i} only if ϕ_i is positive.

Note 2. When ψ is positive, the operator $[\![\psi \mid \cdot]\!]$ is monotonic.

Then by Tarski's Fixpoint Theorem and Note 1(*iii*) we can use the following alternative definition of $[\![\nu X.\psi]\!]$ in terms of post-fixpoints:

$$[\![\nu X.\psi]\!] = \bigcup\{E \subseteq \Omega \mid E \subseteq [\![\psi \mid E]\!]\}.$$

Let us mention some properties the language $\mathcal{L}_\nu$ can express. First notice that common belief is definable in $\mathcal{L}_\nu$ using the νX operator. An analogous characterization of common knowledge is in [9, Section 11.5].

Note 3. Let ψ be a formula of $\mathcal{L}$. Then $[\![\nu X.\Box(X \wedge \psi)]\!]$ is the event that the event $[\![\psi]\!]$ is common belief.

From now on we abbreviate the formula $\nu X.\Box(X \wedge \psi)$ with ψ a formula of $\mathcal{L}$ to $\Box^*\psi$. So $\mathcal{L}_\nu$ can define common belief. Moreover, as the following observation shows, it can also define the iterated elimination of non-optimal strategies.

Note 4. In the game determined by the event $[\![\nu X.O_\phi X]\!]$, every player selects a strategy which survives the iterated elimination of non-ϕ-optimal strategies.

Proof. It follows immediately from the following equivalence, which is obtained by unpacking the relevant definitions:

$$G_{[\![O_\phi X \wedge X \mid E]\!]} = O_\phi(G_E).$$

5 Proof Systems

Consider the following formula:

$$(rat_\phi \wedge \Box^* rat_\phi) \to \nu X.O_\phi X. \tag{1}$$

By Notes 3 and 4, we see that (1) states that: true common belief that the players are ϕ-rational entails that each player selects a strategy that survives the iterated elimination of non-ϕ-optimal strategies.

In the rest of this section we will discuss a simple proof system in which we can derive (1). We will use an axiom and rule of inference for the fixpoint operator taken from [11] and one axiom for rationality analogous to the one called in [7] an "implicit definition" of rationality. We give these in Figure 3, where, crucially, *ψ is positive in X, and all the ϕ_i's are positive*. We denote here by $\psi[X \mapsto \chi]$ the formula obtained from ψ by substituting each occurrence of the variable X with the formula χ. Assuming given some standard proof rules for propositional reasoning, we add the axioms and rule given in Figure 3 to obtain the system **P**.

Axiom schemata

$rat_\phi \to (\Box\chi \to O_\phi\chi)$ $\quad ratDis$

$\nu X.\psi \to \psi[X \mapsto \nu X.\psi]$ νDis

Rule of inference

$$\frac{\chi \to \psi[X \mapsto \chi]}{\chi \to \nu X.\psi}\ \nu Ind$$

Fig. 3. Proof system **P**

A formula is a ***theorem*** of a proof system if it is derivable from the axioms and rules of inference. An $\mathcal{L}_\nu$-formula ψ is ***valid*** if for every belief model $(\Omega, \ldots)$ for G we have $[\![\psi]\!] = \Omega$. We now establish the soundness of the proof system **P**, that is, that its theorems are valid.

Lemma 2. *The proof system **P** is sound.*

Proof. We show the validity of the axiom $ratDis$:

Let $(\Omega, \overline{s}_1, \ldots, \overline{s}_n, P_i, \ldots, P_n)$ be a belief model for G. We must show that $[\![rat_\phi \to (\Box\chi \to O_\phi\chi)]\!] = \Omega$. That is, that for any χ the inclusion $[\![rat_\phi]\!] \cap [\![\Box\chi]\!] \subseteq [\![O_\phi\chi]\!]$ holds. So take some $\omega \in [\![rat_\phi]\!] \cap [\![\Box\chi]\!]$. Then for every $i \in [1..n]$, $\phi_i(\overline{s}_i(\omega), G_{P_i(\omega)})$, and $P_i(\omega) \subseteq [\![\chi]\!]$. So by monotonicity of ϕ_i, $\phi_i(\overline{s}_i(\omega), G_{[\![\chi]\!]})$, i.e. $\omega \in [\![O_{\phi_i}\chi]\!]$ as required.

The axioms νDis and the rule νInd were introduced in [11]; they formalise, respectively, the following two consequences of Tarski's Fixpoint Theorem concerning a monotonic operator F:

– νF is a post-fixpoint of F, i.e., $\nu F \subseteq F(\nu F)$ holds,
– if Y is a post-fixpoint of F, i.e., $Y \subseteq F(Y)$, then $Y \subseteq \nu F$.

Next, we establish the already announced claim.

Theorem 1. *The formula (1) is a theorem of the proof system* $\boldsymbol{P}$.

Proof. The following formulas are instances of the axioms $ratDis$ (with $\psi := \Box^* rat_\phi \wedge rat_\phi$) and νDis (with $\psi := \Box(X \wedge rat_\phi)$) respectively:

$$rat_\phi \to (\Box(\Box^* rat_\phi \wedge rat_\phi) \to O_\phi(\Box^* rat_\phi \wedge rat_\phi)), \tag{2}$$
$$\Box^* rat_\phi \to \Box((\Box^* rat_\phi) \wedge rat_\phi). \tag{3}$$

Putting these two together via some propositional logic, we obtain

$$((\Box^* rat_\phi) \wedge rat_\phi) \to O_\phi((\Box^* rat_\phi) \wedge rat_\phi),$$

which is of the right shape to apply the rule νInd (with $\chi := \Box^* rat_\phi \wedge rat_\phi$ and $\psi := O_\phi X$). We then obtain

$$(\Box^* rat_\phi \wedge rat_\phi) \to \nu X.O_\phi X,$$

which is precisely the formula (1).

Corollary 1. *The formula (1) is valid.*

It is interesting to note that no axioms or rules for the modalities $\Box$ or O were needed in order to derive (1), other than those connecting them with rationality. In particular, no introspection is required on the part of the players, nor indeed is the K axiom $\Box(\varphi \wedge \psi) \leftrightarrow (\Box\varphi \wedge \Box\psi)$ needed.

In the language $\mathcal{L}_\nu$, the rat_{ϕ_i} are in effect propositional constants. We might instead define them in terms of the $\Box_i$ and O_{ϕ_i} modalities but to this end we would need to extend the language $\mathcal{L}_\nu$. One way to do this is to use a quantified modal language, allowing quantifiers over set variables, so extending $\mathcal{L}_\nu$ by allowing formulas of the form $\forall X\varphi$. Such quantified modal logics are studied in [10]. It is straightforward to extend the semantics to this larger class of formulas:

$$[\![\forall X\varphi \mid E]\!] = \{\omega \in \Omega \mid \forall F \subseteq \Omega, \omega \in [\![\varphi \mid F]\!]\}.$$

In the resulting language each rat_{ϕ_i} constant is definable by a formula of this second-order language:

$$rat_{\phi_i} \equiv \forall X(\Box_i X \to O_{\phi_i} X). \tag{4}$$

The following observation then shows correctness of this definition.

Note 5. For all $i \in [1..n]$ the formula (4) is valid in the semantics sketched.

To complete our proof-theoretic analysis we augment the proof system **P** with the following proof rule where we assume that χ is positive in X, but where ψ is an arbitrary ν-free $\mathcal{L}_\nu$-formula:

$$\frac{\chi \to \psi}{\nu X.\chi \to \nu X.\psi}\ Incl$$

The soundness of this rule is a direct consequence of Lemma 1.

To formalize the statement **Imp** we need two optimality conditions, gbr_i and lsd_i.

To link the proof systems for the languages $\mathcal{L}_O$ and $\mathcal{L}_\nu$ we add the following proof rule, where each ϕ_i and ψ_i is an optimality condition in $\mathcal{L}_O$, and $O_\phi X \to O_\psi X$ is a formula of $\mathcal{L}_\nu$.

$$\frac{\phi_i \to \psi_i, i \in [1..n]}{O_\phi X \to O_\psi X}\ Link$$

The soundness of this rule is a direct consequence of the semantics of the formulas $O_\phi X$ and $O_\psi X$.

We denote the system obtained from **P** by adding to it the above two proof rules and standard first-order logic rules concerning the formulas in the language $\mathcal{L}_O$, like

$$\frac{\exists y\, \forall x \phi}{\forall x\, \exists y \phi}$$

by **R**. We can now formalize the statement **Imp** as follows:

$$(rat_{gbr} \wedge \Box^* rat_{gbr}) \to \nu x.O_{lsd}x. \tag{5}$$

The following result then shows that this formula can be formally derived in the considered proof system.

Theorem 2. *The formula (5) is a theorem of the proof system **R**.*

Proof. The properties gbr_i are monotonic, so the following implication is an instance of (1):

$$(rat_{gbr} \wedge \Box^* rat_{gbr}) \to \nu x.O_{gbr}x.$$

Further, since the implication $gbr_i \to lsd_i$ holds, we get by the *Link* rule

$$\nu x.O_{gbr}x \to \nu x.O_{lsd}x,$$

from which (5) follows.

Corollary 2. *The formula (5) is valid.*

6 Summary

We have studied the logical form of epistemic characterisation results, for arbitrary (including infinite) strategic games, of the form "common knowledge of ϕ-rationality entails playing according to the iterated elimination of non-ϕ' properties". A main contribution of this work is in revealing, by giving syntactic proofs, the reasoning principles involved in two cases: firstly when $\phi = \phi'$ (Theorem 1), and secondly when ϕ entails ϕ' (Theorem 2). In each case the result holds when ϕ is monotonic. The language $\mathcal{L}_\nu$ that we used to formalise this reasoning is to our knowledge novel in combining optimality operators with fixpoint notions. Such a combination is natural when studying such characterisation results, since common knowledge and iterated elimination are both fixpoint notions.

The language $\mathcal{L}_\nu$ is parametric in the optimality conditions used by players. It is therefore built on the top of a first-order language $\mathcal{L}_O$ used to define syntactically optimality conditions relevant for our analysis.

References

1. Apt, K.R.: Relative strength of strategy elimination procedures. Economics Bulletin 3(21), 1–9 (2007), available from `http://economicsbulletin.vanderbilt.edu/Abstract.asp?PaperID=EB-07C70015`
2. Apt, K.R.: The many faces of rationalizability. Berkeley Electronic Journal of Theoretical Economics 7(1) (2007), 38 pages
3. Aumann, R.J., Brandenburger, A.: Epistemic conditions for nash equilibrium. Econometrica 63(5), 1161–1180 (1995)
4. Battigalli, P., Bonanno, G.: Recent results on belief, knowledge and the epistemic foundations of game theory. Research in Economics 53, 149–225 (1999)
5. Benthem, J.v.: Rational dynamics and epistemic logic in games. International Game Theory Review 9(1), 13–45 (2007), (Erratum reprint, 9(2), 377–409)
6. Bernheim, B.D.: Rationalizable strategic behavior. Econometrica 52, 1007–1028 (1984)
7. Bruin, B.d.: Explaining Games: On the logic of game theoretic explanations. Ph.D. thesis, ILLC, Amsterdam (2004)
8. Dawar, A., Grädel, E., Kreutzer, S.: Inflationary fixed points in modal logics. ACM Transactions on Computational Logic (TOCL 5(2), 282 – 315 (2004)
9. Fagin, R., Halpern, J.Y., Vardi, M., Moses, Y.: Reasoning about knowledge. MIT Press, Cambridge, MA (1995)
10. Fine, K.: Propositional quantifiers in modal logic. Theoria 36, 336–346 (1970)
11. Kozen, D.: Results on the propositional mu-calculus. Theoretical Computer Science 27(3), 333–354 (1983)
12. Lipman, B.L.: A note on the implications of common knowledge of rationality. Games and Economic Behaviour 6, 114–129 (1994)
13. Osborne, M.J., Rubinstein, A.: A Course in Game Theory. MIT Press, Cambridge, MA (1994)
14. Tarski, A.: A lattice-theoretical fixpoint theorem and its applications. Pacific Journal of Mathematics 5, 285–309 (1955)